Niebel, Thomas; Rasel, Fabienne; Viete, Steffen

Conference Paper

BIG Data - BIG Gains? Understanding the Link Between Big Data Analytics and Innovation

28th European Regional Conference of the International Telecommunications Society (ITS): "Competition and Regulation in the Information Age", Passau, Germany, July 30 - August 2, 2017

Provided in Cooperation with:

International Telecommunications Society (ITS)

BIG Data - BIG Gains? Understanding the Link Between Big Data Analytics and Innovation

Thomas Niebel* Fabienne Rasel † Steffen Viete‡

June 30, 2017

Work in Progress

Abstract

This paper analyzes the relationship between firms' use of big data analytics and their innovative performance for product innovations. Since big data technologies provide new data information practices, they create new decision-making possibilities, which firms can use to realize innovations. Applying German firm-level data we find suggestive evidence that big data analytics matters for the likelihood of becoming a product innovator as well as the market success of the firms' product innovations. The regression analysis reveals that firms which make use of big data have a higher likelihood of realizing product innovations as well as a higher innovation intensity. Interestingly, the results are of equal magnitude in the manufacturing and services industries. The results support the view that big data analytics have the potential to enable innovation.

Keywords: Big data, data-driven decision-making, innovation, product innovation, firm-level data.

JEL Classification Numbers: D22, L20, O33.

*Corresponding author: ZEW Mannheim, *email*: niebel@zew.de, Centre for European Economic Research (ZEW) Mannheim, ICT Research Department, P.O. Box 103443, 68034 Mannheim, Germany.

†ZEW Mannheim, rasel@zew.de

‡ZEW Mannheim, viete@zew.de.

For further information on projects of the authors see www.zew.de/staff_tni, www.zew.de/staff_frl and www.zew.de/staff_sve as well as the ZEW annual report on www.zew.de/en. We thank Irene Bertschek, Chris Forman and participants at the ZEW ICT seminar for helpful comments and suggestions. All remaining errors are ours.

1 Introduction

Latest technological trends like connected devices and machines, wearables, and the universal application of sensors but also (user-generated) online content are drivers of the constantly increasing amount of data. As a reference to the large volumes of diverse data and associated new data information practices that have become available to firms, *big data analytics* has become an important topic among practitioners, policy makers and scientists. Broadly speaking, the concept of big data indicates the amount and complexity of newly available data and the technical challenges to process them (Dumbill, 2013). A more narrow definition of the term, which is commonly used in the literature, highlights the following three characteristics: (1) enormous amount of data (volume), (2) variety of data coming from highly diverse sources (variety), and (3) the pace of data processing (velocity). Enormous progress in computing power, storage capacity, and software have been necessary for the surge of big data technologies.

Much of the debate and research has centered around possible implications for firms and businesses. As big data alters the sources and types of information available to decision-makers in the firm, it is expected to impact on established ways of decision- and strategy-making, which traditionally relied on predefined data collected for specific needs (Constantiou and Kallinikos, 2015). In particular, data which has become available to firms is often not collected intentionally, but in a heterogenous and unstructured way (Varian, 2010; Anderson, 2008). The ability to analyze such data, extract insights and appropriate value from it presents one of the key challenges for firms. One problem big data poses to decision-making is that correlations identified from the raw data are erroneously interpreted as causal relationships or that misleading patterns are found in the data (McAfee and Brynjolfsson 2012). Starting from such data patterns found with big bata analytics, decisions without potential for improvement or even wrong decisions can be made. That is why the use of big data analytics may not guarantee sustainable, positive effects on firm performance (‚Big Gains‘). The vague situation with respect to privacy, data protection, the regulatory environment, or insufficient internet connection are viewed as other main barriers for the diffusion of big data.

Despite these challenges associated with big data, a widely shared expectation is that the ongoing changes in how data is being generated and made relevant for firms can help to increase business value by using data profitably, which sometimes even used to be produced as ‚waste‘ product of the business activity before the surge of big data technologies. New data information practices and better informed decision-making can be particularly advantageous for firms' innovation processes, which often involve high uncertainty and risk. In this vein, mining of consumption patterns and consumer sentiment analysis, for instance, might improve the adoption and market success of new products. Data obtained from sensors can facilitate the detection of product defects and the subsequent improvement of existing products. Insights obtained from big data can furthermore reduce the duration and costs of the innovation process. Besides improving the R&D process, big data can also be at the core of the innovation itself. Monitoring transactions and

combining different information facilitates the development of new personalized services (Varian, 2010) and other data intensive innovations. Consequently, big data is expected to enable firms from all industries to create new products and services, improve existing ones, and to develop new business models (McKinsey Global Institute, 2011).

High potentials to foster innovation, productivity, and growth are also ascribed to big data by policymakers. For instance, the European Commission (EC) stresses the importance of data for growth and innovation in a knowledge-based economy in their policy report on the strategy for a digital single market. Furthermore, the EC has already taken measures to promote the data-driven economy, e.g. through public-private-partnerships for projects on big data or by supporting the development of standards and interoperability in data usage (European Commission, 2014).

Despite the high expectations associated with big data and the prominent position it takes as a current key technological trend, there is little empirical evidence on its effect on firm performance overall, and firms' innovation performance in particular. Against this background, we analyze the relation of firms' use of big data and innovation performance using large scale firm-survey data from German manufacturing and services industries. Extending classical knowledge production functions by firms' use of big data, we find that big data information practices are associated with a higher propensity to innovate, as well as higher innovation intensity.

Our paper contributes to the literature in various respects: (i) we provide first large scale empirical evidence based on representative firm-level data on the role of big data for firm performance in terms of product innovation activities of manufacturing and service firms. (ii) The paper further contributes to a better understanding on the relationship between data analysis and innovation output across industries and helps to assess the potential benefits of big data analytics.

The remainder of the article is structured as follows. Section 2 reviews the empirical literature on the potential effects of big data analytics on firm performance. Section 3 lays out our empirical framework. Section 4 describes the data and measures. Section 5 and 6 discuss the descriptive and econometric results. Finally, section 7 concludes.

2 Related Empirical Literature

The reports of McKinsey Global Institute (2011) and OECD (2015) provide a general overview of the definition and application scope of big data analytics and the potential economic benefits that may return from the use of big data technologies and of data-driven innovation.[1] Up to now, empirical evidence on the potential effects of big data analytics on firm performance is scarce. There only exist few empirical studies based on selective U.S. datasets for specific sectors or limited to listed companies (e.g., (Brynjolfsson et al., 2011); (Tambe, 2014); (Brynjolfsson and McElheran, 2016)). The common finding of those studies is that firms with more intensive data

[1] Goodridge and Haskel (2015) develop an economic framework to determine the importance of big data on GDP and on GDP growth. Applying their framework to the UK, they find that big data in form of transformed data and data-based knowledge accounted for 0.02 per cent of growth in market sector value added from 2005 to 2012.

usage are more productive. Furthermore, some studies show complementarities between big data usage and highly qualified employment (e.g., (Tambe, 2014); (Brynjolfsson and McElheran, 2016)).

Concerning the diffusion process of data-related activities, Saunders and Tambe (2015) demonstrate an increasing trend of the use of data-related activities in U.S. firms within the IT industry in the period from 1996 to 2012. Likewise, Brynjolfsson and McElheran (2016) find that the use of data-driven decision-making almost tripled in the U.S. during the period from 2005 to 2010, whereas the adoption was particularly high in larger firms and in firms with more skilled workers and a higher IT capital stock.

With respect to the role of data-driven decision-making for productivity, Brynjolfsson et al. (2011) find that such practices are related with a 5 to 6 per cent increase in productivity and output among publicly traded U.S. firms. Similarly, Brynjolfsson and McElheran (2016) show that data-related management practices caused a productivity increase of 3 per cent for firms in the U.S. manufacturing sector. However, the authors highlight heterogeneity in the productivity returns of data-related practices according to firm characteristics: The productivity return of data-related management practices seems to be lower for larger, older and capital-intensive multi-unit firms. In addition, they find evidence for complementarity between data-driven decision-making and a high IT capital stock prior to the adoption of data-related practices as well as complementarity between data practices and the presence of better-educated workers.

Tambe (2014) shows evidence for labor market complementarities between investments in and productivity returns from a particular big data technology, namely Hadoop, and the availability of employees with the skills for using this big data technology. The hypotheses for labor market complementarities between technology and human capital are supported by findings that U.S. firms' Hadoop investments yield higher productivity returns in geographic labor markets with high availability of workers with Hadoop skills. Wu and Hitt (2016) find evidence for complementarity between data analysis skills and process-related decisions, which is suggested by positive productivity returns for firms with a higher level of employees' data skills and the use of practices that aim at improving business processes.

Overall, the findings of the role of big data analytics for firm performance are compatible with prior evidence on complementarity and performance effects of ICT. There is a large literature on the productivity effects of ICT investment as well as on complementarities between ICT and human capital.[2] Generally, ICT is viewed as an enabler for innovation (e.g., (Brynjolfsson and Saunders, 2010), (Spiezia, 2011)). In terms of the role of data use for realizing innovation, Bertschek and Kesler (2017) find that the adoption of a Facebook page and the user activity on this page are significant determinants for the realization of a product innovation.

To the best of our knowledge, there is no study yet that examines explicitly the role of big data analytics for innovation performance at the firm level across industries. Based on the findings from the literature on the role of big data for firm performance and generally the contribution of ICT to

[2]For an overview see e.g., Draca et al. (2007), Van Reenen et al. (2010), Cardona et al. (2013).

innovation, we expect a positive relationship between big data analytics and product innovation - however, possibly not uniformly for all firms but contingent on potential complementary factors.

3 Empirical Framework

We analyze the contribution of big data to firms' innovation performance within the widely used knowledge production function framework introduced by Griliches (1979). This framework postulates a transformation process which links various inputs associated with knowledge accumulation, such as investments in R&D or human capital, to the firms' innovative output. Knowledge production functions have been the workhorse model in understanding the importance of various knowledge sources besides formal R&D. In the present work, we explicitly account for big data in the firms' knowledge production processes in order to provide first insights into the relevance of big data for firms' innovation activities.

The following section outlines our empirical model of the knowledge production function. We denote y_{1i}^* the latent propensity of firm i to achieve product innovations, given the firm's use of big data analytics, $bigdata_i$, as well as the firm's R&D intensity and other firm- and market-specific characteristics denoted by the vector $\mathbf{c}_{1i}$. For simplicity of the formal exposition of the analysis, let us further collect the variable on the firm's big data use and further control variables in the vector $\mathbf{x}_1 \equiv (bigdata, \mathbf{c}_1)$. The first step of the empirical model of the knowledge production function assumes a linear additive relationship and amounts to

$$y_{1i}^* = \beta_1 bigdata_i + \gamma_1' \mathbf{c}_{1i} + \epsilon_{1i} = \delta_1' \mathbf{x}_{1i} + \epsilon_{1i} \tag{1}$$

where β denotes the parameter of interest, capturing the effect of the firm's engagement in big data analytics on the propensity to innovate. ϵ_{1i} denotes an idiosyncratic error term, which captures unobserved variables affecting y_{1i}^* and is assumed to be identically and independently normally distributed, $\epsilon_{1i} \sim NID(0, \sigma_1^2)$. The observed variable is the innovation success, i.e. the event of introducing a new product to the market, y_{1i}, which is defined by the following observation rule:

$$y_{1i} = \mathbf{1}[y_{1i}^* > 0] \tag{2}$$

where $\mathbf{1}[\,]$ is the indicator function taking the value 1 if the condition is satisfied and 0 otherwise. Equations (1) and (2) describe the first part of our analysis, in which we estimate the relationship between the use of big data and firms' innovation propensity via a simple Probit model.[3]

Beyond the relationship between big data and the propensity to innovate, we want to asses the relationship with the firms' innovation intensities. Thus, let y_{2i}^* denote the firms' potential innovation intensities given the firm's use of big data, R&D intensity and further firm- and market-

[3]Given the distributional assumption in Equation (1), we have $P(y_{1i} = 1|\mathbf{x}_{1i}) = P(y_{1i}^* > 0|\mathbf{x}_{1i}) = P(\epsilon_{1i} \leq \mathbf{x}_{1i}'\beta) = \Phi(\mathbf{x}_{1i}'\beta)$ under the normalization restriction $\sigma_1^2 = 1$, which we estimate by Maximum Likelihood.

specific characteristics, such that

$$y_{2i}^* = \beta_2 bigdata_i + \gamma_2' \mathbf{c}_{2i} + \epsilon_{2i} = \delta_2' \mathbf{x}_{2i} + \epsilon_{2i} \tag{3}$$

where, again, $\epsilon_{2i} \sim NID(0, \sigma_2^2)$ denotes the normally distributed idiosyncratic error term and $\mathbf{x}_2 \equiv (bigdata, \mathbf{c}_2)$. In line with much of the empirical literature studying innovation intensities, the observed innovation intensity, which is typically measured by the sales ratio of innovative products and services, is assumed to be defined by the following observation rule:

$$y_{2i} = \mathbf{1}[y_{2i}^* > 0] y_{2i}^*. \tag{4}$$

Equations (3) and (4) together results in the standard Tobit model (Tobin, 1958), which takes account of the nonlinear nature of the conditional expectation function $E(y_{2i}|\mathbf{x}_{2i})$ due to the nontrivial fraction of firms which do not generate sales with newly introduced products.[4]

The conditional expectation for the model made up of Equations (3) and (4) is given by

$$E(y_{2i}|\mathbf{x}_{2i}) = \Phi(\delta_2' \mathbf{x}_{2i}/\sigma)\delta_2' \mathbf{x}_{2i} + \sigma\phi(\delta_2' \mathbf{x}_{2i}/\sigma) \tag{5}$$

where $\Phi_i(\cdot)$ and $\phi_i(\cdot)$ denote the standard normal cumulative distribution function and density function, respectively.[5]

A potential problem in estimating the Tobit model arises due to its strong and restrictive distributional assumptions. Unlike Ordinary Least Squares estimation, in cases of heteroskedasticity or non-normality, Tobit estimates will generally be inconsistent.[6] Due to the limitations of the standard Tobit model, we check our results against the fractional logit model proposed by Papke and Wooldridge (1996). This model builds on the logistic distribution function to model the conditional expectation of a fractional dependent variable

$$E(y_{2i}|\mathbf{x}_{2i}) = \frac{exp(\delta_2' \mathbf{x}_{2i})}{1 + exp(\delta_2' \mathbf{x}_{2i})}. \tag{6}$$

Using a Bernoulli link function the model is estimated by Maximum Likelihood. Importantly for our application, the fractional logit model allows for y_{2i} to take on the boundaries 0 and 1 with positive probability, as opposed to other common solutions to model proportions, such as using the logit transformation of y_{2i} (e.g. Mohnen et al., 2006; Raymond et al., 2015).

The standard Tobit and the fractional logit model discussed above assumes that the observed innovation intensity is the result of a single process influenced by the same set of determinants. As

[4]Note that, in line with the general literature, in the Tobit model with zero lower limit we ignore the upper limit of the innovation intensity. However, as the share of observations at the upper limit (of 1) is well below 1%, we regard the effect of upper limiting cases on the estimates to be negligible.

[5]For a more detailed description of Tobit type models see for instance Amemiya (1984) or Maddala (1986).

[6]Note that the assumption of normality and constant variance of ϵ_{2i} is crucial in deriving the conditional expectation in Equation (5).

the innovation intensity is a fractional variable with a lot of observations clustering at zero, one possible concern is that a single model fitted to all data might be insufficient. In particular, while big data might be related to the propensity to innovate, it could at the same time be unrelated to the innovation intensity, i.e. the market success of the firms' innovations, conditional on being an innovator. In that case, the simple Tobit model in Equations (3) and (4) is too restrictive. Alternatively, we can consider a framework in which the models for the propensity to innovate and for the innovation intensity conditional on being an innovator differ. Overall, there is no consensus in the empirical innovation literature whether a one part model, such as the simple Tobit model described above, or an alternative two step model is more appropriate to model firms' innovation intensities.[7] We therefore also estimate an alternative two step model and test the one against the other. In particular, we consider that, alternative to Equation (4), the observed innovation intensity is defined by the observation rule

$$y_{2i} = \mathbf{1}[y_{1i}^* > 0] y_{2i}^* \tag{7}$$

such that the sales ratio of innovations is observed if the firm's propensity to innovate is sufficiently large. In addition, let the unobserved errors $(\epsilon_{1i}, \epsilon_{2i})$ be jointly normally distributed with covariance σ_{12}. Equations (3) and (7) together with the distributional assumptions on the error terms yield the Tobit Type II or Heckman Selection model, in which the conditional expectations of interest are given by:

$$E(y_{1i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}) = \Phi(\delta_1' \mathbf{x}_{1i}) \tag{8}$$

$$E(y_{2i}|\mathbf{x}_{1i}, \mathbf{x}_{2i}, y_{1i} = 1) = \delta_2' \mathbf{x}_{2i} + \sigma_{12} \frac{\phi(\delta_1' \mathbf{x}_{1i})}{\Phi(\delta_1' \mathbf{x}_{1i})} \tag{9}$$

Given both models, the simple Tobit as well as the Heckman Selection model, are being used in the empirical innovation literature, we estimate both to check the robustness of our findings to the common modeling assumptions.

As a main caveat, our study is subject to common endogeneity concerns in the empirical literature on the value of ICT. Omitted variables might confound the relation between the use of big data and firms' innovation performance. The main advantage of our data is the wide variety of background characteristics we can account for. In particular, our data contain rich information on firms' use of alternative digital technologies, which help to disentangle the quality and features of big data analytics activities from the firms' general ICT intensities as well as the use of legacy systems. Since the empirical literature on ICT performance generally suffers from a lack of good instrumental variables, reverse causation is another common endogeneity concern. We note that our study faces the risk of being confounded by reverse causation since we are only able to provide

[7] See for instance Cassiman and Veugelers (2006), Andries and Czarnitzki (2014) or Hottenrott and Lopes-Bento (2016) for other studies applying both types of models to model innovation shares.

controlled correlation applying a new cross-sectional dataset. Nevertheless, we believe that our analysis is an important first step in understanding how firms make use of big data analytics and to shed light on the often claimed role of big data technologies in fostering innovation in adopting firms.

4 Data and Measures

Our analysis is based on the ZEW ICT survey which is a survey of manufacturing and services firms located in Germany with five or more employees.[8] In total, six waves were collected in 2000, 2002, 2004, 2007, 2010 and 2015. We exploit the wave of 2015, which contains information on the firms' use of big data for the first time. About 4400 firms were interviewed about their characteristics and particularly about their ICT usage. The data were collected via computer-aided telephone interviews (CATI) based on a sample stratified with respect to industry and firm size. The respondent is usually from the board of management or as fall back option the head of the IT department.[9]

4.1 Big Data Analytics

Our main variable of interest is the dummy variable for big data analytics that is equal to one in case the firm is using big data technologies. More precisely, the following question was asked in our survey:

> *"Up next a question about so called big data, i.e. the processing of large amounts of data. Does your company systematically analyze large amounts of data to support business operations?"*

As we aim at measuring firms' engagement with big data across different industries and firm sizes, our measure of big data use leaves room for subjective assessment of the interviewee. This was done deliberately, because despite the public recognition of big data as one of the current key technologies, the term lacks a generic definition and does not constitute a unified concept. The most commonly accepted definition is based on the "3 Vs" . They are the enormous amount of data (volume), (2) the variety of data coming from highly diverse sources (variety), and (3) the pace of data processing (velocity). Furthermore, the technology for big data has been advancing quickly over time. As the size of datasets is continuously increasing and tools that are more sophisticated arise to analyze them, big data has always been an evolving concept. The definition of big data might also be contingent on the industrial context and depend on the specific software used and the common size of datasets in a particular industry (McKinsey Global Institute, 2011). Product innovations based on big data analytics will also vary a lot between industries. For

[8]The data are available at the ZEW Research Data Centre - http://kooperationen.zew.de/en/zew-fdz.
[9]For more information about the survey see Bertschek et al. (2017).

instance, Luckow et al. (2015) describe potential innovations in the automotive industry. Based on the steadily number of sensors per vehicle, new innovative services like traffic prediction, safety warnings, vehicle diagnostics, and location-based services are based on big data analytics. Another example is that insurance companies make use of different data sources and big data technologies to design improved premium policies.

4.2 Innovation Outcomes

Our data include items on innovation and R&D activities following the Community Innovation Survey (CIS) and the guidelines of the Oslo Manual by the OECD and Eurostat (Mortensen et al., 2005). In particular, we consider the event of introducing a product innovation to the market as the first outcome of the knowledge production process. The relevant measure is a binary indicator, which takes the value one if the firm has introduced a new or substantially improved product or service to the market over the past three years (*Product Innovation*). The product can be new to the market overall or new to the firm. In addition to the propensity to innovate, we investigate the intensity of innovation, which we measure by the share in total sales due to new products in the year 2013 (*% of Sales New Product*). Our innovation intensity measure captures the market success of product innovations (Mairesse and Mohnen, 2002; Laursen and Salter, 2006).

4.3 Control Variables

Following the empirical innovation literature, we control for an extensive set of firm characteristics which have been shown to affect innovation performance. We measure R&D intensity, the potentially single most important input factor to knowledge production, as R&D expenditures over total sales (*% of R&D Expenses*). The firms' R&D intensities affects the propensity to innovate as well as the firms' innovation successes (Pakes and Griliches, 1980) and reflect the relative importance of innovation activities for the firm. As firms, which are making use of big data analytics are likely to be generally more intensive ICT users and have high ICT intensities, in turn, ICT intensity can be expected to positively affect firms' innovation performance (Hempell and Zwick, 2008). Therefore, we control for firms' ICT intensities by the share of employees mainly working with personal computers (*% of Emp. Predom. Using PC*) as well as the share of employees having access to the internet at the workplace (*% of Emp. Using Internet*). Furthermore, as the use of enterprise software systems has been shown to be related to firms'innovation activities (Engelstätter, 2012), we include a binary variable into the model indicating whether or not the firm has an enterprise software system implemented (*Enterprise Software*). We note that our additional measures on the firms' ICT use capture the effect of mature software systems and data technologies, which lack the quality of large-scale data analytics, such as SAP and other standard Enterprise Resource Planning Systems or conventional databases based on Structured Query Language. Furthermore, firms' innovative capabilities are affected by the employees' human capital, their knowledge, abi-

lities and creativity (Vinding, 2006). Thus, we control for the share of highly skilled employees, i.e. workers with degrees from universities and technical colleges (*% Highly Qualified Employees*), as well as the share of employees with vocational training (*% Medium Qualified Employees*). We furthermore account for the age structure of the workforce by controlling for the share of employees below 30 years of age (*% of Employees < Age 30*) and above 50 years of age (*% of Employees > Age 50*). As the maturity of the firm might affect both, the use of cutting edge technology as well as their innovative capabilities (Huergo and Jaumandreu, 2004), we control for the years since the founding year of the firm (*Age*). Younger firms might furthermore achieve higher sales shares with new products merely because they have less established products in their portfolio. Firm size has been found to be important for technology adoption (Haller and Siedschlag, 2011). Likewise, potential relations between firm size and innovation have already been found by Schumpeter (1942). Overall, larger firms can be expected to have better internal financial resources and enjoy economies of scale and scope, which benefits both, technology adoption as well as innovative capabilities. We thus control for firm size measured by the log of the number of employees (*Employees*). As the likelihood of innovating has been shown by some studies to increase with physical capital intensity (e.g. Lööf and Heshmati, 2006), we control for the log of gross investments (*Investment*). The exposure to international product markets affects the potential market size for new products as well as the competitive pressure to innovate (Hottenrott and Lopes-Bento, 2016). We thus include an indicators whether the firm exports to foreign markets (*Exporter*) and whether it is part of a multinational enterprise (*Multinational*). We additionally account for the firms' ownership structure by a binary variable indicating whether the firm is part of a national enterprise group (*Group*). Finally, we account for structural regional differences between the two former German states by a binary indicator for location in former Eastern Germany (*East Germany*) as well as structural differences between industries by including a set of 16 industry dummies constructed from 3-digit NACE industry codes.[10]

5 Descriptive Statistics

Table 5.1 provides summary statistics on the variables used in the analysis. The share of firms having introduced new products or services amounts to 48 per cent and the average share of sales due to new products and services is 8.4 per cent. In our estimation sample, 22 per cent of the firms are relying on big data to support their decision making. With a share of 56 per cent considerably more firms have an enterprise software system implemented. About 45 per cent of the employees predominately work with computers. The average number of employees in the sample is 94, so the sample mainly consists of small and medium-sized enterprises.

[10]Table A.1 provides an overview over the industries and their distribution in the estimation sample.

Table 5.1: Summary Statistics: Estimation Sample

	N	Mean	Median	SD	Min	Max
Product Innovation	2727	0.48	0	0.50	0	1
% of Sales New Product	2727	0.084	0	0.15	0	1
Big Data	2727	0.22	0	0.41	0	1
% of Emp. Predom. Using PC	2727	0.45	0.33	0.34	0	1
% of Emp. Using Internet	2727	0.57	0.50	0.37	0	1
Enterprise Software	2727	0.56	1	0.50	0	1
% of R&D Expenses	2727	0.051	0.0059	0.11	0	1
Employees	2727	93.6	25	263.2	5	4500
Employees (in logs)	2727	3.44	3.22	1.31	1.61	8.41
Investment in Mill. Euro	2727	0.91	0.10	4.68	0.00050	130
Investment (in logs)	2727	-2.02	-2.30	1.84	-7.60	4.87
Exporter	2727	0.45	0	0.50	0	1
% Highly Qualified Employees	2727	0.19	0.10	0.24	0	1
% Medium Qualified Employees	2727	0.63	0.70	0.27	0	1
% of Employees < Age 30	2727	0.24	0.20	0.17	0	1
% of Employees > Age 50	2727	0.27	0.25	0.19	0	1
East Germany	2727	0.24	0	0.43	0	1
Age (in logs)	2727	3.17	3.14	0.92	0	6.39
Group	2727	0.30	0	0.46	0	1
Multinational	2727	0.095	0	0.29	0	1

We apply the data to shed light on the incidence of data driven decision making and on the question which firms exploit data strategically for their decision making. Figure A.1 provides the in sample share of firms which are using big data analytics by industry. Overall, the use of data analytics is higher in the services sector. Data driven decision making has proliferated in the financial sector, where over half of the firms in the sample indicate to systematically apply data for strategic support of their business operations. Firms in the retail and wholesale trade sectors are also intensive in data use for their decision process with a diffusion of around 30 percent. Amongst the manufacturing industries, big data is used most intensively in the chemicals and motor vehicles sectors, by around 23 percent of the firms in each of the two sectors. The sector in which least firms rely on data for their decision making is manufacturing of durable goods with a diffusion rate of only 13 percent. Figure A.1 additionally depicts the share of firms innovating by industry. Among manufacturer of chemicals, electronics and machinery as well as in the ICT services sector over 70 per cent of firms introduced new products or services within the previous three years. The share of innovating firms is lowest in the transport services sector with only 23 per cent. Overall, the variation over industries depicted in Figure A.1 does not provide a clear picture on the relation between the use of big data and innovation performance. While some sectors with a high diffusion of big data also exhibit high shares of innovating firms, this is certainly not true for all industries. For example, while in the manufacturing of machinery industry around 70 per cent of the firms innovate, only 16 per cent rely on big data for their decision making.

To further investigate which firms exploit data strategically for their decision making, Table A.2 provides summary statistics of firm characteristics conditional on the firms' use of big data.

Overall, firms which have introduced big data technologies are using ICT more intensively overall, are larger in terms of employees and investments, have higher R&D expenditures, more often belong to a multi plant or multinational firm and are more often exporting. Importantly, firms using big data analytics are on average more innovative, both at the extensive and intensive margin. Still, a thorough investigation of the relation of big data with firms' innovation performance calls for a multivariate analysis as outlined above.

6 Econometric Results

The following section provides the main estimation results. Table 6.1 presents the estimation results of the Probit models analyzing the relation between big data utilization and the firms' innovation propensity for the full sample as well as for the estimation sample split into the manufacturing and services sector, respectively. The estimate of the coefficient on the big data indicator is positive and statistically significant in all three estimations. Moreover, the estimated relation between big data use and the likelihood to introduce a new product or service to the market is economically meaningful. Looking at the results for the full sample in column (1), the firms' application of big data analytics is associated with a 6.7 percentage point increase in the propensity to innovate. Interestingly, the results are of comparable magnitude when differentiating between manufacturing and service firms in columns (2) and (3). The respective results show that firms with big data analytics in use are 6.6 percentage points more likely to innovate in the manufacturing sector and 6.8 percentage points more so in the services sector. Looking at the estimated coefficients on other control variables, in particular those for other measures of ICT use by the firm, we find that the firms' general ICT intensity measured by the share of employees working predominantly with PCs is positively and significantly related to the innovation propensity only among firms of the service sector. Our estimation results furthermore confirm existing research on the positive relation between enterprise software and innovation (e.g. Engelstätter, 2012). ERP Systems typically serve for the planning and controlling of business processes across different parts of the value chain. They moreover constitute a platform to integrate more specific applications, such as Supply Chain Management or Customer Relationship Management Software. While firms which are using ERP-Systems are typically integrating information across different business processes and engage in data driven decision-making, the features of classical ERP Software systems lack the quality of big data analytics in terms of amount of data that is being processed and the software tools which are used to analyze the data. Importantly, our measure for big data use explains the firms' innovation propensity beyond the effect of these legacy software systems. Further strong predictors for the firms' likelihood to innovate over all three models are the firms' R&D intensity, the firms' export status as well as the indicator whether or not the firm belongs to a multinational enterprise.

Table 6.2 reports the results from the Tobit and the Fractional Logit estimations modelling the

Table 6.1: Dependent Variable: Dummy for Product Innovation - Probit Regression - Average Marginal Effects

	(1) Full Sample	(2) Manufacturing	(3) Services
Big Data	0.067***	0.066*	0.068**
	(0.023)	(0.035)	(0.029)
% of Emp. Predom. Using PC	-0.000	-0.089	0.058
	(0.042)	(0.074)	(0.051)
% of Emp. Using Internet	0.080**	0.080	0.074
	(0.035)	(0.049)	(0.052)
Enterprise Software	0.086***	0.115***	0.064**
	(0.020)	(0.030)	(0.026)
% of R&D Expenses	0.912***	1.123***	0.776***
	(0.159)	(0.263)	(0.178)
Employees (in logs)	0.010	0.015	0.009
	(0.012)	(0.017)	(0.015)
Investment (in logs)	0.024***	0.019*	0.029***
	(0.007)	(0.011)	(0.010)
Exporter	0.165***	0.145***	0.183***
	(0.021)	(0.029)	(0.032)
% Highly Qualified Employees	0.159***	0.372***	0.043
	(0.061)	(0.123)	(0.079)
% Medium Qualified Employees	-0.040	-0.017	-0.097
	(0.043)	(0.055)	(0.069)
% of Employees < Age 30	-0.026	-0.068	-0.002
	(0.052)	(0.076)	(0.071)
% of Employees > Age 50	-0.022	-0.054	0.009
	(0.049)	(0.069)	(0.070)
East Germany	0.005	0.031	-0.037
	(0.021)	(0.028)	(0.030)
Age (in logs)	-0.008	0.005	-0.023
	(0.010)	(0.014)	(0.014)
Group	0.035*	0.054*	0.015
	(0.020)	(0.030)	(0.028)
Multinational	0.134***	0.132***	0.122**
	(0.035)	(0.046)	(0.054)
Industry Dummies	Yes	Yes	Yes
Pseudo R^2	0.207	0.182	0.212
Observations	2727	1415	1312
Log likelihood	-1496.289	-794.057	-692.757

Standard errors in parentheses, * p<0.10, ** p<0.05, *** p<0.01.

All models include an intercept.

sales share of new products, i.e. the market success of the firms' innovations. The table reports average marginal effects on the conditional expectations in Equations (5) and (6). Overall, results show that the use of big data is not only related to the firms' innovation status, but also to the firms' innovation intensity. Over both empirical models in all three samples, big data is positively and statistically significantly associated with the sales share of innovations. Again the estimates are economically meaningful and of equal magnitude for the full sample and within the manufacturing and the services sector. In particular, for the full sample (columns (1) and (2)) the use of big data is associated with a 2.5 to 2.9 percentage point increase in the sales share from innovations. All other coefficients are in line with prior expectations. R&D intensity is a strong predictor of the sales share of innovations. Over most specifications, the firms' age is negatively associated with innovation intensity. Thus, younger firms achieve more sales with newly introduced products or services.

Finally, we turn to the estimation results of the Heckman Selection Model. Theoretically, the model is identified by the functional form assumptions. That is, even if the set of regressors in both equations of the model is identical ($\mathbf{x}_1 = \mathbf{x}_2$), the model is identified due to the nonlinearity of the inverse Mills ratio in the second equation.[11] However, in practice it is desirable to have an exclusion restriction, i.e. a variable that enters the selection equation but not the second equation, for more reliable identification of the model parameters (e.g. Little and Rubin, 2014). Ideally, the exclusion restriction is selected on theoretical grounds. However, there is no variable available which theoretically affects the firms' likelihood to innovate while leaving the firms' innovation intensity unaffected. We thus follow, for instance, Andries and Czarnitzki (2014) or Peters and Schmiele (2010) and search for an exclusion restriction empirically in order to ensure that identification of the model parameters does not merely rest on functional form assumptions. When including the full set of variables in both equations of the model, the firms' export status is strongly and significantly related to the firms' propensity to innovate, whereas the respective parameter estimate in the second equation is very small and statistically insignificant (see Table A.3 in the appendix for the respective estimation results). We thus rely on the firms' export status as an exclusion restriction. We note, however, that the validity of our exclusion restriction cannot be tested.

Table 6.3 reports the average marginal effects of the Heckman model estimation. For each of the three samples, the first column reports the partial effects on the propensity to innovate while the second column reports the expected innovation intensity, conditional on being an innovator, according to Equation (9). Overall, the previous results are confirmed by the estimation of the selection model. The application of big data analytics is associated with a 6.6 percentage point higher innovation propensity over all samples. The estimated partial effect on the innovation intensity conditional on being an innovator ranges between 2.2 percentage points in the full sample and 2.6 percentage points in manufacturing sector. Note that, in contrast, the use of enterprise

[11] The inverse Mills ratio corresponds to the term $\frac{\phi(\delta_1' \mathbf{x}_{1i})}{\Phi(\delta_1' \mathbf{x}_{1i})}$ in Equation (9).

Table 6.2: Dependent Variable: % Share of New Products in Turnover- Tobit/FracReg Regressions

	Full Sample		Manufacturing		Services	
	(1)	(2)	(3)	(4)	(5)	(6)
	Tobit	FracReg	Tobit	FracReg	Tobit	FracReg
Big Data	0.025***	0.029***	0.027***	0.033***	0.025***	0.028***
	(0.006)	(0.008)	(0.009)	(0.011)	(0.008)	(0.010)
% of Emp. Predom. Using PC	0.006	0.009	-0.006	-0.000	0.018	0.022
	(0.011)	(0.013)	(0.019)	(0.022)	(0.014)	(0.015)
% of Emp. Using Internet	0.018*	0.016	0.022*	0.023	0.015	0.008
	(0.010)	(0.011)	(0.013)	(0.015)	(0.014)	(0.018)
Enterprise Software	0.020***	0.018***	0.031***	0.029***	0.013**	0.012
	(0.005)	(0.006)	(0.007)	(0.008)	(0.007)	(0.008)
% of R&D Expenses	0.253***	0.195***	0.319***	0.240***	0.199***	0.157***
	(0.020)	(0.024)	(0.035)	(0.049)	(0.023)	(0.024)
Employees (in logs)	-0.007**	-0.014***	-0.004	-0.010*	-0.009**	-0.019***
	(0.003)	(0.004)	(0.005)	(0.005)	(0.004)	(0.005)
Investment (in logs)	0.007***	0.008***	0.003	0.002	0.011***	0.014***
	(0.002)	(0.003)	(0.003)	(0.004)	(0.003)	(0.004)
Exporter	0.037***	0.030***	0.034***	0.028***	0.039***	0.030***
	(0.005)	(0.007)	(0.007)	(0.009)	(0.008)	(0.010)
% Highly Qualified Employees	0.036**	0.027	0.055**	0.025	0.016	0.028
	(0.016)	(0.019)	(0.028)	(0.032)	(0.022)	(0.023)
% Medium Qualified Employees	-0.015	-0.018	-0.019	-0.033	-0.019	-0.001
	(0.012)	(0.015)	(0.015)	(0.020)	(0.019)	(0.021)
% of Employees < Age 30	0.003	0.015	0.015	0.035	-0.008	0.002
	(0.014)	(0.017)	(0.020)	(0.023)	(0.018)	(0.023)
% of Employees > Age 50	-0.002	-0.001	-0.009	-0.001	0.004	0.001
	(0.013)	(0.015)	(0.018)	(0.023)	(0.018)	(0.020)
East Germany	0.002	0.003	0.012	0.013	-0.009	-0.008
	(0.006)	(0.006)	(0.008)	(0.009)	(0.008)	(0.009)
Age (in logs)	-0.008***	-0.011***	-0.004	-0.008*	-0.011***	-0.015***
	(0.003)	(0.003)	(0.004)	(0.004)	(0.004)	(0.005)
Group	0.008	0.008	0.015*	0.016	0.001	-0.001
	(0.005)	(0.007)	(0.008)	(0.010)	(0.007)	(0.009)
Multinational	0.024***	0.023**	0.015	0.011	0.034**	0.037**
	(0.009)	(0.009)	(0.011)	(0.011)	(0.014)	(0.016)
Industry Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.363	0.092	0.403	0.069	0.330	0.126
Observations	2727	2727	1415	1415	1312	1312
Censored	1441		636		805	
Uncensored	1286		779		507	
Log likelihood	-663.292	-715.540	-255.512	-413.578	-383.393	-298.865

software is only positively statistically significantly related to the propensity to innovate, while the estimated partial effect on the conditional innovation intensity is negative, small and statistically insignificant.

Finally, note that over all three models we cannot reject independence between the two equations. Consequently, we can re-estimate the equation modeling the firms' innovation intensity on the subsample of innovating companies only. In fact, all results from above were confirmed and detailed regression results are thus omitted for the sake of brevity.

Table 6.3: Heckman Selection Model with exclusion restriction, Marginal Effects

	Full Sample		Manufacturing		Services	
	(1)	(2)	(3)	(4)	(5)	(6)
	1st	2nd	1st	2nd	1st	2nd
Big Data	0.066***	0.022***	0.066*	0.026**	0.066**	0.023*
	(0.022)	(0.008)	(0.034)	(0.010)	(0.029)	(0.013)
% of Emp. Predom. Using PC	-0.002	0.002	-0.094	0.016	0.056	-0.008
	(0.043)	(0.017)	(0.071)	(0.022)	(0.053)	(0.027)
% of Emp. Using Internet	0.080**	-0.004	0.079	0.009	0.076	-0.026
	(0.036)	(0.014)	(0.050)	(0.016)	(0.053)	(0.028)
Enterprise Software	0.086***	-0.007	0.115***	-0.003	0.064**	-0.011
	(0.020)	(0.008)	(0.030)	(0.010)	(0.026)	(0.013)
% of R&D Expenses	0.950***	0.221***	1.230***	0.282***	0.796***	0.179***
	(0.112)	(0.025)	(0.202)	(0.039)	(0.128)	(0.037)
Employees (in logs)	0.011	-0.020***	0.016	-0.011**	0.009	-0.031***
	(0.011)	(0.004)	(0.017)	(0.006)	(0.015)	(0.007)
Investment (in logs)	0.024***	0.004	0.018*	-0.004	0.029***	0.014***
	(0.008)	(0.003)	(0.011)	(0.004)	(0.010)	(0.005)
% Highly Qualified Employees	0.160***	0.003	0.378***	-0.041	0.041	0.046
	(0.061)	(0.023)	(0.113)	(0.034)	(0.081)	(0.040)
% Medium Qualified Employees	-0.037	-0.011	-0.015	-0.027	-0.097	0.034
	(0.043)	(0.017)	(0.056)	(0.019)	(0.069)	(0.037)
% of Employees < Age 30	-0.025	0.031	-0.068	0.069***	-0.003	-0.000
	(0.051)	(0.020)	(0.075)	(0.026)	(0.069)	(0.032)
% of Employees > Age 50	-0.022	0.007	-0.051	0.005	0.005	0.016
	(0.048)	(0.019)	(0.067)	(0.023)	(0.069)	(0.035)
East Germany	0.004	0.002	0.031	0.005	-0.038	0.003
	(0.021)	(0.008)	(0.028)	(0.009)	(0.030)	(0.015)
Age (in logs)	-0.008	-0.013***	0.006	-0.009*	-0.023	-0.019***
	(0.010)	(0.004)	(0.013)	(0.005)	(0.014)	(0.007)
Group	0.034*	-0.000	0.051*	0.008	0.015	-0.011
	(0.021)	(0.007)	(0.031)	(0.009)	(0.028)	(0.013)
Multinational	0.132***	0.011	0.128***	-0.005	0.122**	0.031
	(0.035)	(0.010)	(0.046)	(0.011)	(0.054)	(0.019)
Exporter	0.163***	0.000	0.139***	0.000	0.181***	0.000
	(0.021)	(.)	(0.029)	(.)	(0.031)	(.)
Industry Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2727	2727	1415	1415	1312	1312
$\hat{\sigma}_{12}$	-0.238		-0.282		-0.291	
LR-Test $H_0 : \sigma_{12} = 0$ [$\chi^2(1)$], p-Val	0.127		0.113		0.203	
Log Likelihood	-977.274		-411.139		-533.017	

7 Preliminary Conclusions and Future Research

This paper investigates the association between the use of big data analytics and firms' propensity to innovate, as well as firms' innovation performance, which we measure by the sales share due to new products or services and which constitutes a measure of the market success of the firms' innovations. Our preliminary results show that the use of big data analytics is associated with a higher propensity to innovate, as well as higher innovation performance. Importantly, this relation holds once we control for the use of mature software systems and data technologies, such as Enterprise Resource Planning Software, which lack the quality of big data analytics. These results are robust with respect to various alternative specifications and econometric methods used (Tobit,

Fractional Regressions, Heckman Selection Models). As the knowledge production process and innovative output likely differ between manufacturing and service firms, we investigate potential effect heterogeneity with regard to the two sectors. Interestingly, the associations we measure are of similar magnitude among firms in the manufacturing and the service industry. Overall, our results are consistent with positive returns of big data analytics on product innovations at the extensive and intensive margin. Moreover, our findings suggest that big data analytics have the potential to support innovative activity and they support the view that data is a valuable input to the production process. In our ongoing research, we will further investigate heterogenous effects of big data, with regard to firm characteristics highlighted in the existing literature, such as general ICT intensity, human capital, or firm size. Overall, our findings constitute an important first step towards a better understanding of the value of large scale data to support the firm's innovation activities.

References

Amemiya, T. (1984), 'Tobit Models: A Survey', *Journal of Econometrics* **24**(1-2), 3–61.

Anderson, C. (2008), 'The end of theory: The data deluge makes the scientific method obsolete', *Wired magazine* **16**(7), 16–07.

Andries, P. and Czarnitzki, D. (2014), 'Small Firm Innovation Performance and Employee Involvement', *Small Business Economics* **43**(1), 21–38.

Bertschek, I. and Kesler, R. (2017), Let the User Speak: Is Feedback on Facebook a Source of Firms' Innovation?, ZEW Discussion Paper 17-015, Center for European Economic Research.

Bertschek, I., Ohnemus, J. and Viete, S. (2017), 'The ZEW ICT Survey 2002 to 2015: Measuring the Digital Transformation in German Firms', *Jahrbücher für Nationalökonomie und Statistik* .

Brynjolfsson, E., Hitt, L. M. and Kim, H. H. (2011), 'Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?'. Available at SSRN: http://dx.doi.org/10.2139/ssrn.1819486.

Brynjolfsson, E. and McElheran, K. (2016), Data in Action: Data-Driven Decision Making in U.S. Manufacturing, Working Papers 16-06, Center for Economic Studies, U.S. Census Bureau.

Brynjolfsson, E. and Saunders, A. (2010), *Wired for Innovation: How IT is Reshaping the Economy*, The MIT Press.

Cardona, M., Kretschmer, T. and Strobel, T. (2013), 'ICT and Productivity: Conclusions from the Empirical Literature', *Information Economics and Policy* **25**(3), 109–125.

Cassiman, B. and Veugelers, R. (2006), 'In Search of Complementarity in Innovation Strategy: Internal R&D and External Knowledge Acquisition', *Management science* **52**(1), 68–82.

Constantiou, I. D. and Kallinikos, J. (2015), 'New Games, New Rules: Big Data and the Changing Context of Strategy', *Journal of Information Technology* **30**(1), 44–57.

Draca, M., Sadun, R. and Van Reenen, J. (2007), Productivity and ICT: A Review of the Evidence, *in* C. Avgerou, R. Mansell, D. Quah and R. Silverstone, eds, 'The Oxford Handbook of Information and Communication Technologies', Oxford University Press, pp. 100–147.

Dumbill, E. (2013), 'Making Sense of Big Data', *Big Data* **1**(1), 1–2.

Engelstätter, B. (2012), 'It's not all About Performance Gains – Enterprise Software and Innovations', *Economics of Innovation and New Technology* **21**(3), 223–245.

European Commission (2014), 'Towards a Thriving Data-Driven Economy'. COM(2014) 442 final. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1404888011738&uri=CELEX:52014DC0442. Last Accessed: 26 March 2017.

Goodridge, P. and Haskel, J. (2015), How Does Big Data Affect GDP? Theory and Evidence for the UK, Discussion Paper 2015/06, Imperial College Business School.

Griliches, Z. (1979), 'Issues in Assessing the Contribution of R&D to Productivity Growth', *Bell Journal of Economics* **10**(1), 92–116.

Haller, S. A. and Siedschlag, I. (2011), 'Determinants of ICT Adoption: Evidence from Firm-Level Data', *Applied Economics* **43**(26), 3775–3788.

Hempell, T. and Zwick, T. (2008), 'New Technology, Work Organisation, And Innovation', *Economics of Innovation and New Technology* **17**(4), 331–354.

Hottenrott, H. and Lopes-Bento, C. (2016), 'R&D Partnerships and Innovation Performance: Can There be too Much of a Good Thing?', *Journal of Product Innovation Management* **33**(6), 773–794.

Huergo, E. and Jaumandreu, J. (2004), 'How Does Probability of Innovation Change with Firm Age?', *Small Business Economics* **22**(3-4), 193–207.

Laursen, K. and Salter, A. (2006), 'Open for Innovation: The Role of Openness in Explaining Innovation Performance among UK Manufacturing Firms', *Strategic Management Journal* **27**(2), 131–150.

Little, R. J. and Rubin, D. B. (2014), *Statistical Analysis with Missing Data*, Wiley.

Lööf, H. and Heshmati, A. (2006), 'On the Relationship Between Innovation and Performance: A Sensitivity Analysis', *Economics of Innovation and New Technology* **15**(4-5), 317–344.

Luckow, A., Kennedy, K., Manhardt, F., Djerekarov, E., Vorster, B. and Apon, A. (2015), Automotive Big Data: Applications, Workloads and Infrastructures, *in* '2015 IEEE International Conference on Big Data (Big Data)', IEEE, pp. 1201–1210.

Maddala, G. S. (1986), *Limited-Dependent and Qualitative Variables in Econometrics*, number 3 *in* 'Econometric Society Monographs', Cambridge University Press.

Mairesse, J. and Mohnen, P. (2002), 'Accounting for Innovation and Measuring Innovativeness: An Illustrative Framework and an Application', *The American Economic Review* **92**(2), 226–230.

McAfee, A. and Brynjolfsson, E. (2012), 'Big Data: The Management Revolution', *Harvard Business Review* **90**(10), 60–68.

McKinsey Global Institute (2011), 'Big Data: The Next Frontier for Innovation, Competition, and Productivity'. Available at http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation. Last Accessed: 9 February 2016.

Mohnen, P., Mairesse, J. and Dagenais, M. (2006), 'Innovativity: A Comparison Across Seven European Countries', *Economics of Innovation and New Technology* **15**(4-5), 391–413.

Mortensen, P. S., Bloch, C. W. et al. (2005), *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data*, Organisation for Economic Cooperation and Development, OECD, Paris.

OECD (2015), 'Data-Driven Innovation: Big Data for Growth and Well-Being'.

Pakes, A. and Griliches, Z. (1980), 'Patents and R&D at the Firm Level: A First Report', *Economics Letters* **5**(4), 377–381.

Papke, L. E. and Wooldridge, J. M. (1996), 'Econometric Methods for Fractional Response Variables with an Application to 401 (k) Plan Participation Rates', *Journal of Applied Econometrics* **11**(6), 619–632.

Peters, B. and Schmiele, A. (2010), The Influence of International Dispersed vs. Home-Based R&D on Innovation Performance, ZEW Discussion Paper 10-102, Center for European Economic Research.

Raymond, W., Mairesse, J., Mohnen, P. and Palm, F. (2015), 'Dynamic Models of R&D, Innovation and Productivity: Panel Data Evidence for Dutch and French Manufacturing', *European Economic Review* **78**, 285–306.

Saunders, A. and Tambe, P. (2015), 'Data Assets and Industry Competition: Evidence from 10-K Filings'. Available at SSRN: http://dx.doi.org/10.2139/ssrn.2537089.

Schumpeter, J. A. (1942), *Capitalism, Socialism and Democracy*, Harper & Brothers, London.

Spiezia, V. (2011), 'Are ICT Users More Innovative?: an Analysis of ICT-Enabled Innovation in OECD Firms', *OECD Journal: Economic Studies* **2011**(1), 1–21.

Tambe, P. (2014), 'Big Data Investment, Skills, and Firm Value', *Management Science* **60**(6), 1452–1469.

Tobin, J. (1958), 'Estimation of relationships for limited dependent variables', *Econometrica: Journal of the Econometric Society* pp. 24–36.

Van Reenen, J., Bloom, N., Draca, M., Kretschmer, T. and Sadun, R. (2010), *The Economic Impact of ICT*, Centre for Economic Performance, London School of Economics. Available at http://ec.europa.eu/information_society/newsroom/cf/dae/document.cfm?doc_id=669. Last Accessed: 9 February 2016.

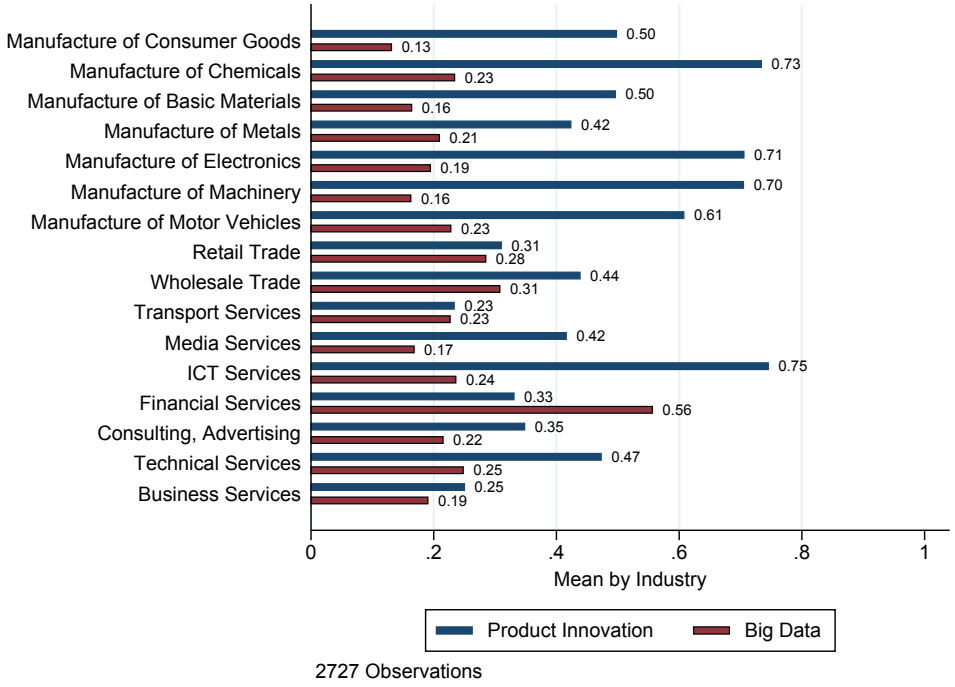
Varian, H. R. (2010), 'Computer Mediated Transactions', *The American Economic Review* **100**(2), 1–10.

Vinding, A. L. (2006), 'Absorptive Capacity and Innovative Performance: A Human Capital Approach', *Economics of Innovation and New Technology* **15**(4-5), 507–517.

Wu, L. and Hitt, L. M. (2016), How Do Data Skills Affect Firm Productivity: Evidence from Process-driven vs. Innovation-driven Practices. Working Paper,.

A Appendix

Figure A.1: Industry Means of Product Innovation and Big Data: Estimation Sample



Source: ZEW ICT survey 2015.

Table A.1: Distribution of Firms across Industries: Estimation Sample

	N	Percentage
Manufacture of Consumer Goods	450	16.50
Manufacture of Chemicals	94	3.45
Manufacture of Basic Materials	250	9.17
Manufacture of Metals	196	7.19
Manufacture of Electronics	180	6.60
Manufacture of Machinery	166	6.09
Manufacture of Motor Vehicles	79	2.90
Retail Trade	158	5.79
Wholesale Trade	130	4.77
Transport Services	150	5.50
Media Services	125	4.58
ICT Services	161	5.90
Financial Services	133	4.88
Consulting, Advertising	158	5.79
Technical Services	129	4.73
Business Services	168	6.16
Total	2727	100.00

Table A.2: Summary Statistics by Big Data Use of Firms: Estimation Sample

	not		Big Data		Total	
	N	Mean	N	Mean	N	Mean
Product Innovation	2134	0.45	593	0.60	2727	0.48
% of Sales New Product	2134	0.07	593	0.12	2727	0.08
Big Data	2134	0.00	593	1.00	2727	0.22
% of Emp. Predom. Using PC	2134	0.42	593	0.55	2727	0.45
% of Emp. Using Internet	2134	0.55	593	0.65	2727	0.57
Enterprise Software	2134	0.51	593	0.78	2727	0.56
% of R&D Expenses	2134	0.04	593	0.07	2727	0.05
Employees	2134	65.73	593	193.88	2727	93.60
Employees (in logs)	2134	3.24	593	4.18	2727	3.44
Investment in Mill. Euro	2134	0.57	593	2.16	2727	0.91
Investment (in logs)	2134	-2.28	593	-1.07	2727	-2.02
Exporter	2134	0.44	593	0.49	2727	0.45
% Highly Qualified Employees	2134	0.19	593	0.21	2727	0.19
% Medium Qualified Employees	2134	0.63	593	0.61	2727	0.63
% of Employees < Age 30	2134	0.23	593	0.26	2727	0.24
% of Employees > Age 50	2134	0.28	593	0.26	2727	0.27
East Germany	2134	0.25	593	0.22	2727	0.24
Age (in logs)	2134	3.13	593	3.30	2727	3.17
Group	2134	0.26	593	0.43	2727	0.30
Multinational	2134	0.08	593	0.15	2727	0.09

Table A.3: Heckman Selection Model (no exclusion restriction), Marginal Effects

	Full Sample		Manufacturing		Services	
	(1)	(2)	(3)	(4)	(5)	(6)
	1st	2nd	1st	2nd	1st	2nd
Big Data	0.065***	0.022***	0.066*	0.025**	0.066**	0.023*
	(0.022)	(0.008)	(0.034)	(0.010)	(0.029)	(0.013)
% of Emp. Predom. Using PC	-0.003	0.002	-0.095	0.018	0.056	-0.010
	(0.043)	(0.017)	(0.071)	(0.023)	(0.053)	(0.028)
% of Emp. Using Internet	0.079**	-0.004	0.079	0.009	0.076	-0.026
	(0.036)	(0.014)	(0.050)	(0.016)	(0.053)	(0.028)
Enterprise Software	0.086***	-0.007	0.115***	-0.003	0.064**	-0.013
	(0.020)	(0.008)	(0.030)	(0.010)	(0.026)	(0.013)
% of R&D Expenses	0.956***	0.221***	1.244***	0.282***	0.800***	0.178***
	(0.111)	(0.026)	(0.201)	(0.039)	(0.127)	(0.038)
Employees (in logs)	0.010	-0.020***	0.015	-0.011**	0.009	-0.031***
	(0.011)	(0.004)	(0.017)	(0.006)	(0.015)	(0.008)
Investment (in logs)	0.024***	0.004	0.018	-0.004	0.029***	0.014**
	(0.008)	(0.003)	(0.011)	(0.004)	(0.010)	(0.005)
Exporter	0.165***	-0.007	0.142***	-0.010	0.184***	-0.010
	(0.021)	(0.008)	(0.029)	(0.011)	(0.031)	(0.014)
% Highly Qualified Employees	0.160***	0.002	0.378***	-0.044	0.041	0.048
	(0.061)	(0.023)	(0.113)	(0.034)	(0.081)	(0.041)
% Medium Qualified Employees	-0.037	-0.011	-0.015	-0.028	-0.097	0.036
	(0.043)	(0.017)	(0.056)	(0.019)	(0.069)	(0.038)
% of Employees < Age 30	-0.025	0.031	-0.067	0.068***	-0.003	-0.001
	(0.051)	(0.020)	(0.075)	(0.026)	(0.069)	(0.033)
% of Employees > Age 50	-0.022	0.007	-0.050	0.006	0.004	0.015
	(0.048)	(0.020)	(0.067)	(0.023)	(0.069)	(0.036)
East Germany	0.004	0.001	0.031	0.004	-0.037	0.002
	(0.021)	(0.008)	(0.028)	(0.009)	(0.030)	(0.016)
Age (in logs)	-0.008	-0.013***	0.006	-0.009*	-0.023	-0.019***
	(0.010)	(0.004)	(0.013)	(0.005)	(0.014)	(0.007)
Group	0.034*	-0.001	0.051*	0.007	0.015	-0.012
	(0.021)	(0.008)	(0.030)	(0.010)	(0.028)	(0.013)
Multinational	0.132***	0.011	0.127***	-0.005	0.121**	0.033*
	(0.035)	(0.010)	(0.046)	(0.011)	(0.054)	(0.020)
Industry Dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2727	2727	1415	1415	1312	1312
$\hat{\sigma}_{12}$	-0.273		-0.316		-0.340	
LR-Test $H_0 : \sigma_{12} = 0$ $[\chi^2(1)]$, p-Val	0.089		0.067		0.143	
Log Likelihood	-976.942		-410.712		-532.725	